DOCUMENT RESUME

ED 218 305 TM 820 339

AUTHOR Jones-Loheyde, Katherine; And Others
TITLE Evaluation of Student Writing: Comparison of Holistic Grades and Scores by Trained Raters on "Guide to Analysis of Written Discourse."
PUB DATE 20 Mar 82
NOTE 34p.; Paper presented at the Annual Meeting of the American Educational Research Association (66th, New York, NY, March 19-23, 1982).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Analysis; Evaluation Methods; Expository Writing; Higher Education; *Holistic Evaluation; *Scoring; *Writing (Composition); *Writing Evaluation; Writing Skills

ABSTRACT

Two methods of evaluating student writing were compared. Holistic evaluation attempts to assess the overall quality of a composition. It is a guided procedure for sorting or ranking written composition. A rater scores the composition by placing it within the range of papers produced in response to a given assignment. Scoring may also be done using an analytic rating scale. For this study, students were assigned a composition based on data provided thereby prompting analytic writing. Students' papers were holistically scored by their instructors. Trained raters also scored the papers using an analytical rating system, Guide to the Analysis of Written Discourse (GAWD). The relationship between the ratings derived on GAWD and the holistic grades for each composition was statistically analyzed. Comparison of the two methods of evaluation indicated the holistically derived grade was a more reliable measure of writing ability than the analytic scale. There was a correlation between both methods of evaluation. Both were useful for measuring change in writing ability. (DWH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED218305

Evaluation of Student Writing: Comparison of Holistic Grades and Scores by Trained Raters on "Guide to Analysis of Written Discourse"

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Katherine Jones-Loheyde
Karen Jambeck
Jorge Esquilin

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

K. Jones-Loheyde

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 820 339

A paper presented at the American Educational Research Association annual meeting, New York, New York; March 20, 1982.

List of Figures and Tables

Page

Figure:

1. Summary of Data Collected 7

2. A Guide to the Analysis of Written Discourse. . . . 8

Table:

1. Repeated Measures ANOVA for the Total Ratings in September and December. 12

2. Frequency Table of Holistically Derived Grades Over Time By Section. 13

3. Comparison of Holistic Mean Scores Over Time. . . . 14

4. Results of the Tuckey (HDS) Method Applied to the Holistically Derived Grade Means. 15

5. Repeated Measures ANOVA for Holistic Evaluation . . 16

6. Scores on Components of Rating Scale. 18

7. Correlation Coefficients Between Rating Scale Components 20

8. Shared Variance Between Instruments 22

Evaluation of Student Writing: Comparison of Holistic Grades and Scores by Trained Raters on "Guide to Analysis of Written Discourse"

> "The only undisputable fact that has emerged from the continuing controversy over students' writing is that the evaluation of writing has become big business."[1]

Introduction

Evaluation of student writing has proved--and continues to prove--an issue of controversy and disagreement to those of us concerned with the teaching of and research in composition. Admittedly, writing assessment is an area of testing fraught with difficulties. Unlike testing for content or testing for demonstrated proficiency in an area like music or sport, writing does not clearly lend itself to clearly defined measureable qualities. Little agreement is found between raters in assessing writing and little agreement is found as to what constitutes good or even reasonable levels of proficiency in writing. Standards are nebulous, purposes vary and measurement approaches vary with those purposes.

That a range of methodologies for evaluating writing have been spawned in the past few years is readily apparent. Central though is the point that each method rests on its own set of assumptions rather than evolving from commonly held assumptions or beliefs about evaluation.[2] Statistical compilations of reader responses provide the rational for holistic scoring and analytic scales; developmental stages of language acquisition account for quantification of syntactic features; and a triangler model of discourse underlies primary trait scoring.

Odell and Cooper raise three questions of concern to teachers, researchers, and specialists in writing evaluation:

1. what assumptions are implicit or explicit in our evaluation procedures?;
2. are those assumptions consistent with current discourse theory?;
3. will the results of using these procedures help us with the problem of improving students writing?.[3]

In considering these questions, Odell and Cooper examined four different approaches to evaluation: The General Impression scoring procedure used by the Educational Testing Service (ETS) in scoring the English Composition Test, one of the achievement test series of the College Entrance Examination Board;[4] the analytic scale developed by Paul D. Diederich;[5] and assessment of "relative readability" proposed by E. D. Hirsch, Jr.;[6] and the Primary Trait scoring procedure developed by Richard Lloyd-Jones for the National Assessment of Educational Progress.[7]

As a corollary to the premise that different demands made of writers elicit different performances from them, Richard Lloyd-Jones[8] contends that writing must not be judged by general criteria for good writing but by what Lloyd-Jones calls "primary traits"--criteria that are uniquely suited to the specific task a writer is trying to perform. The guides used in such scoring represent a significant departure from the customary practice of judging all expository writing by a single set of criteria, e. g., the categories found in an analytic scale.[9]

Of critical consideration in the study that follows are these questions raised by Cooper: would different evaluation procedures lead us to make different judgments about a given student's writing performance? Is any one evaluation procedure especially helpful or harmful to students or to teachers?[10]

Of specific concern here is the comparison of two methods of evaluation: holistic grading and scoring using an analytic rating scale. Holistic evaluation attempts to assess the overall quality of a composition, and although various aspects of content and form may be considered, no marking or tallying of these aspects is required.

Charles R. Cooper[11] defines holistic evaluation as a guided procedure for sorting or ranking written composition. Raters respond to a written sample by either a) matching it with another sample in a graded series of papers; b) scoring the sample for the prominence of certain features inherent in that kind of writing; or c) assigning the sample a letter or number grade. He cites seven examples of holistic evaluation: essay scale, analytic scale, dichotomous scale, feature analysis, primary trait scoring, general impression marking, and what Peter Elbow[12] refers to as the "center of gravity" response. Of these various systems of holistic scoring, general impression marking is the simplest in that it requires no detailed discussion of specific features of the writing under consideration nor a summary of scores given to separate features.[13] A rater scores the composition by placing the paper within the range of papers produced in response to a given assignment.

This procedure is used by the Educational Testing Service and the College Entrance Examination Board to score the English Composition Test[14] and the Advanced Placement Test in English.[15] As with any scoring technique, raters must be carefully trained to ensure consensus over multiple raters. This is generally accomplished by reading and discussing a great deal of written work similar to that to be rated. In some cases, raters rank on a given scale: ETS raters utilize a 1 to 4 scale in scoring the English Composition Test and a 1 to 8 scale in scoring the Advanced Placement Test. Holistic scoring in the first reports of the writing assessment of the National Assessment of Educational Progress was accomplished on a scale of 1 to 8.[16] More recently, Connecticut's statewide testing of ninth graders on composition ability was also holistically scored on a scale of one to eight.[17] Those involved in the Beaverton Oregon Program concluded that holistic ratings, while appropriate for program evaluation are much too broad to be of use to the teachers in diagnosing student writing deficiencies. Primary and secondary trait assessment, on the other hand, proved appropriate for diagnosis of writing deficiencies, measurement of student achievement and program evaluation.[18] The four problems noted by French (1966) apply here as in any test requiring actual writing. Sources of error in the scores can stem from any or any combination of the following:

1. student error, in that a student can do well one day and poorly another day on a similar or at the same writing task,

2. test error, in that any testing situation calls for a sample of the students' behavior. Some students may happen to enjoy the topic assigned while others may find it difficult and unstimulating; this results in error,

3. scale error, in that one reader of a composition may be an easy marker and another a tough marker. To get all readers for an examination to grade papers on the same scale is very difficult,

4. reader disagreement in that even if readers could be persuaded to use the same scale, as they were here, their grades would still not look alike because there is disagreement on what kind of writing is best.[19]

The critical problem of reliability among raters of student writing received early notoriety in the work of Starch and Elliott in 1912.[20] Neckel (1963) lists a number of procedures suggested by Dietrich for increasing reliability. Each of his procedures have been influential in this study. His suggestions include having all students write on the same topic; removing the names from the papers; training the readers in marking practices; using two sets of readers; and averaging the grades of two samples of writing obtained from each student at different sessions.[21]

In addition, the findings of Coffman and Kurfman (1968)--that readers employ a global or holistic method in assigning scores rather than an analytic one--has also been used.[22]

Methods

As part of the regular coursework for English 105, (required freshmen composition), students wrote on a topic which had been mutually agreed upon with the course instructors. Samples of such writing were collected twice during the semester, once during the first few weeks (in September) and once again during the last two weeks (December). Thus, from the three sections of English 105, an approximate total of 66 writing samples was collected, as most sections had 22 students enrolled. After discrepancies caused by enrollment changes, complete data were gathered on 61 students.

The Nature of the Assignment

The item of interest here is formal expository prose, specifically that prose produced as a result of analysis. The assignment, therefore, was carefully chosen to prompt analytic writing by providing only data. Both assignments asked the same question, using paragraphs of similar structure, difficulty, and purpose to provide data. A paragraph on women in society by Margaret Mead and "A Proposal to Abolish Grading" by Paul Goodman were used. The question posed required the postulating of a thesis or a conclusion on the part of the writer: "what is the problem? How do you think it can be solved?" (See Appendix A for copies of Assignment I and II.)

Procedures

Each writing sample was graded by the course instructor as the measure of holistic scoring. The original was returned to the student with a copy retained as data for this study.

This enabled the instructor to give one grade for the purpose of the study and another, should he or she wish, to the student. Using the "Guide to Analysis of Written Discourse" (GAWD), established and found reasonably reliable as part of a previous study (Cherkes - Jones Loheyde, 1980),[23] each assignment was rated by two different trained readers.

Each of the three classes provided 22 papers (66 total) twice in the semester for a total of 132 papers. As each paper was rated twice, each reader was asked to code approximately 27 writing samples.

In addition, final course grades of students whose work was evaluated in this study were reported by the instructors. While many different factors effect final course grades, we thought it would be informative to look at changes in writing performance and course grade.

To summarize, the following data was gathered for each student:

Writing Sample	Analytic Scale	Holistic Evaluation
	Content, Organization, Mechanics	
September Paper	Rater I/Rater II; mean score	Grade
December Paper	Rater 1/Rater II; mean score	Grade

Final grade in English 105.

Figure 1.

Thus, questions of interest were as follows:

1. What is the relationship between a score derived from an analytic scale and a grade derived holistically on any one sample of student writing?

2. Is there a significant difference between student writing, as determined by grades and scores, at the beginning and at the end of a given semester?

3. What relationship, if any, exists between two grades derived holistically and the final grade earned in a freshman composition class?

The rating scale, which consisted of eleven items, required trained raters to respond on a scale of one to three: "no evidence" of the feature in the paper, the feature is "implied but not explicit," or the feature is "explicit". Figure 2 shows this instrument.

A Guide to the Analysis of Written Discourse

Organization	No Evidence	Implied But Not Explicit	Explicit
1. There is a thesis apparent.	1	2	3
2. There is one central and controlling idea throughout the essay.	1	2	3
3. The thesis is supported with relevant details beyond simple reiteration of summarization.	1	2	3
4. Logical overall organizations and smooth transitions lead to clear relationship among details and/or ideas.	1	2	3

Content	No Evidence	Implied But Not Explicit	Explicit
5. There is evidence of the formulation of inferences.	1	2	3
6. Individual ideas are expressed with clarity.	1	2	3
Mechanics			
7. There is an absence of distracting errors in word usage such as inappropriate forms of words, "madeup" words, omitted referents, and the like.	1	2	3
8. Each sentence in isolation reads smoothly and clearly.	1	2	3
9. Word choice spelled correctly.	1	2	3
10. Word choice succinct and meaningful.	1	2	3
11. The writer punctuates correctly.	1	2	3
			_____
		Total Score	

Figure 2.

All scoring in papers was done in January. Raters did not know that compositions were written at different times; they only knew that two assignments were addressed.

Scores were determined on the rating scale by both the total and the subscore for each section: organization, content, and mechanics. Keeping in mind that each paper was evaluated by two different readers, scores from the September writing sample ranged from 11 to 33 given the total possible score of 33. The mean total score for the early writing sample--that undertaken in early September was 22.2.

Similarly, for the later paper, written in December, total scores again ranged from 11 to 33, however, the mean was 24.0.

The holistic evaluation was accomplished by instructors in English 105. A grade was assigned to each paper for the purpose of this study. Grades derived holistically by different readers on the September paper ranged from "F" (0.0) to "B" (3.0). The average grade on the fall paper was "D+" (1.6).

The December grades, determined holistically, also ranged from "F" (0.0) to "B" (3.0). However, the average grade on the December paper was 2.1--a little above a "C".

Results

The data were analyzed to determine the relationship between the ratings derived on GAWD and the holistic grades given to each composition. Relationships were tested by computing Pearson product-moment correlation coefficients.

The analytic rating system was further analyzed to examine more closely the subcomponents of the rating system.

1. Relationship between the rating scale scores and the holistic score in September.

The holistic grade given for the September paper relates significantly to the analytic evaluation given by two different raters for the same paper. The Pearson r obtained for the correlation between grade and first rater was .43, while the Pearson r obtained for the correlation between grade and second rater was .33 ($p < .05$).

2. Relationship between rating scale scores and the holistic score in December.

The holistic grade given for the December writing sample bears a low relationship to the ratings given by the two different raters for the same paper. The Pearson r for the correlation between the grade and the first rater in December was .31 ($p < .05$). The Pearson r calculated for the relationship between the grade and the second rater's score was .28 ($p < .05$).

3. Relationship between final course grade in English 105 and rating scale evaluation in September and December.

While many factors are considered in ascertaining final course grades, we also looked at the relationship between the scores on the rating scales across time and the final course grade. A Pearson r was thus calculated for the relationship between the final grade obtained for the course and the total ratings on GAWD in September and December. Each rating, it will be remembered, was derived from calculating the mean of scores from two independent raters.

The relation between final grade and the total ratings in September proved to be low, as would be anticipated, with an r of .30. The relationship between the final grade and the total ratings given in December was moderate with a calculated r of .42. All of these correlations were significant at the .05 level.

4. Changes in ratings over the course of the semester.

A repeated measures ANOVA was performed on the GAWD rating for September and December. The calculated F of 7.99 reveals a significant difference between the total scores on the two papers over time. The mean score on the September paper is 22.38; on the December sample, the mean is 24.14. Using this measure, it can be concluded that there was improvement in writing across time. The figures related to this conclusion are depicted in Table 1.

Repeated Measures ANOVA for the
Total Ratings in September and December

Source	SS	df	MS	F
Between People	1174.79	60	19.58	
Within People	800.32	61	13.12	
Between Measures	94.14	1	94.14	7.99*
Error	706.18	60	11.77	

*significant at the .05 level.

Table 1.

5. Changes in holistically derived scores over the semester.

Using grades assigned by course instructors, the variation in student performance in writing over the course of a semester is shown by section in the chart below. The number of students who received each grade (F through B) at each time is reported. No students received A's on either the September or December writing sample. Grades reported for the samples improved over the time period in which the students were enrolled in freshmen composition.

Frequency Table of
Holistically Derived Grades Over Time
By Section

English 105	Grades in September	Grades in December	Change
Section 2	F 2	1	-1
	D 9	1	-8
	C 9	12	+3
	B 0	6	+6
Total	20	20	
Section 49	F 3	1	-2
	D 7	3	-4
	C 9	12	+3
	B 0	3	+3
Total	19	19	
Section 53	F 0	0	-
	D 2	2	-
	C 14	14	-
	B 6	6	-
Total	22	22	

Table 2.

While students grades on the writing samples rose over the semester in two sections, in one section, no total change was apparent. While grades on individual student's papers may have changed, the total number of students receiving each of the four grades given remained the same in Section 53.

A repeated measures ANOVA was also performed on the data obtained from the holistic measure--the letter grade--given to the 61 students. The grades correspond to evaluation done by the English 105 instructors, obtained twice during the semester (September and December). In addition, a third grade considered was the final grade awarded to the student at the end of the course. This was not, obviously, an average of the previous two grades used in this study, but rather a grade assigned independently.

The mean holistic grade obtained in September is 1.6, in December 2.1, and for the final grade, 2.5. These and the standard deviations are reported in Table 3.

Comparison of Holistic Mean Scores Over Time

Variable	Means	STD Dev	N
Grade 1 (September)	1.64	0.78	61
Grade 2 (December)	2.08	0.69	61
Grade 3 (Final grade)	2.48	0.59	61

Table 3.

Significant differences were found among the means for these three holistic scores. A Tuckey (HDS) method was performed. The results of the procedure are shown in Table 4.

This procedure demonstrates significant differences between the grade in September and the grade given to the December writing sample and between the grade in September and the final grade awarded to students in English 105. In addition, there is a significant difference between the December writing sample grade and the final grade.

Results of the Tuckey (HDS) Method
Applied to the Holistically Derived Grade Means

	$\bar{x}_1$ September Writing Sample	$\bar{x}_2$ December Writing Sample	$\bar{x}_3$ Final Grade
Means	1.64	2.08	2.46
$\bar{X}_i - X_j$		.44	.82
			.38
Q		4.4*	8.2 *
			3.8 *

* $p < .05$ $Q_{cv\ (.05)} = 3.36$ for df = 120

Table 4.

The results from the repeated measures ANOVA show that the differences between the three grades are statistically significant. (See Table 5).

Repeated Measures ANOVA for
Holistic Evaluation

Source	SS	Df	MS	F
Between People	49.67	60	0.83	
Within People	56.66	122	0.46	
Between Measures	20.53	2	10.26	34.10*
Error	36.13	120	0.30	

*significant at the .05<p.

Table 5.

Note that a significant f of 34.10 was calculated.

6. Analysis of components of "Guide to Analysis of Written Discourse" Ratings.

The scores for GAWD are an accumulation of three components parts: organization, content, and mechanics. Scoring on each component is discussed below.

Organization

On the early writing sample collected in September, the mean rating for rater one is 7.9; for rater two 7.4. The total mean rating on organization in September is 7.65. On the second writing sample, written in December at the end of the freshman composition course, the mean rating for the first rater is 8.0 and for the second rater 8.5. Thus, the total mean for organization in December is 8.27. As independently derived scores thus increased from 7.65 to 8.27, we conclude that organizational skill improved during this four month period.

Content

The possible range of scores on content-related questions is two to six. The mean rating for rater one in September is 3.8; the mean rating for rater two on the same factor is 3.6. Thus, the total mean on content on the initial writing sample is 3.7.

On the later sample, the first raters' average score is 4.2; second raters' average is 4.3. The grand mean for content, therefore, is 4.25. Again, we conclude that there was an improvement in content as measured by evidence of formulation of inferences and clear expression of individual ideas across time. All the differences were statistically significant at the $p < .05$.

Mechanics

The mean rating for mechanics on the September writing sample was 10.4, representing the average of rater one's score (10.5) and rater two's score (10.3). On the December writing sample, the total mean was 11.1, representing, again, the average of rater one (10.9) and rater two (11.2). Again, apparent improvement in mechanics is noted.

Table 6 displays the GAWD ratings assigned by each of the two raters on the three component subscores--organization, content, and mechanics--in September and December. The table illustrates growth over time in each of the component elements regardless of rater. The variation in organization and content from September to December

Scores on Components of Rating Scale

	Organization (Range 4 to 12)		Content (Range 2 to 6)		Mechanics (Range 5 to 15)	
	Sept	Dec	Sept	Dec	Sept	Dec
Rater 1 $\bar{x}$	7.9	8.0	3.8	4.2	10.5	10.9
Rater 2 $\bar{x}$	7.4	8.5	3.6	4.3	10.3	11.2
Total $\bar{x}$	7.65	8.27	3.7	4.25	10.4	11.1

Table 6.

was statistically significant. The change in ability in mechanics was not. It is interesting to note on the one hand that the first rater on the three factors in September gave a higher rating than the second rater for the same factor at the same time. On the other hand, in December rater one (whichever of the nine readers was randomly assigned "first" versus "second" status) tended to rate the papers a little lower than rater two.

A correlation matrix among the ratings given for the three subcomponents at different times showed moderate to high correlation coefficients between organization and content (see Table 7). For instance, the correlation coefficient for the first rater between organization and content is .61; for the second rater on the same components,

the correlation coefficients is .60. In December, the correlation coefficient for the first rater on the same items is .51 while that between organization and content for the second rater is .50. Content subscores are also related to mechanics, but the score on mechanics does not correlate with that on organization in either September or December. This is demonstrated in Table 7. It can be concluded that content is highly related to organization and mechanics. This correlation is consistent across time.

Correlation Coefficients Between
Rating Scale Components*

	Organ 1	Organ 2	Organ 3	Organ 4	Cont 1	Cont 2	Cont 3	Cont 4	Mech 1	Mech 2	Mech 3	Mech 4
Organ 1	—											
Organ 2	-0.03	—										
Organ 3	0.09	-0.06	—									
Organ 4	0.02	0.29	0.38	—								
Cont 1	0.61	0.08	0.15	0.03	—							
Cont 2	0.09	0.60	0.05	0.17	0.04	—						
Cont 3	-0.05	0.06	0.51	-0.03	0.07	0.01	—					
Cont 4	0.26	0.13	0.20	0.49	0.17	0.34	0.01	—				
Mech 1	0.17	0.40	0.00	0.15	0.42	0.41	-0.05	0.29	—			
Mech 2	0.20	0.32	0.21	0.28	0.38	0.54	0.06	0.38	0.52	—		
Mech 3	0.16	0.37	0.18	0.01	0.24	0.14	0.61	0.15	0.26	0.14	—	
Mech 4	0.33	0.04	-0.01	0.11	0.05	0.12	-0.03	0.36	0.38	0.31	0.17	—

*1 = Rater 1 in September
2 = Rater 2 in September
3 = Rater 1 in December
4 = Rater 2 in December

*Organ = Organization
Cont = Content
Mech = Mechanics

Table 7.

7. Shared Variance Between Evaluation Instruments

Twelve percent of the variance is accounted for by the holistic grading, while 61% of the variance is accounted for by the analytic subscales added together. When the variance accounted for by both evaluation methods are added together, the total variance accounted for is 73%, which means that only 27% of the improvement in writing ability is due to other factors not considered in this study. Table 8 summarizes this information with the corresponding alpha coefficients.

In comparing the two methods of evaluation, it appears that the holistic derived grade is a more reliable measure of writing ability than the analytic scale, as the calculated alpha for the holistic method amounted to .64, while the alpha for the analytic scale was .40. Moreover, the combined squared multiple r for the holistically derived grades is .61 and the squared multiple r for the analytic scales added together is .12; which means that 73% of the difference in the improvement of writing ability could be accounted for by the use of both methods. Table 8 depicts the values associated with this discussion.

Shared Variance Between Instruments

Variance	(Mult r)^2	Standardized Alpha
1. Holistic Grades	.61	.64
2. Analytic Scale	.12	.40
- total ratings in September	.06	
- total ratings in December	.06	

Table 8.

Discussion

First, there is growth on all measures over time--especially since raters were not aware of the pre- and post-nature of the evaluation. As students improved in all measures over the course of a semester as ganged by both the analytic rating scale and the holistically derived grade, we assume that that improvement may be related to the study of composition. Given that assumption, it appears that the instruction in writing provided in English 105 aided students in addressing issues of organization, content and mechanics.

In comparing the two methods of evaluation, it appears that the holistically derived grade is a more reliable measure of writing ability than is the analytic scale, as the calculated alpha for the grade is .64 as apposed to the alpha for the analytic scale which is .40. Together, however, these two measures account for 73% of the difference of the improvement in writing ability over the course of one semester in freshman English.

Interestingly enough, however, the two measures of evaluation--the holistic and analytic rating scale--are useful for measuring change in writing ability. There is, as well, a significant correlation between both methods of evaluation. Thus, while either instrument--the holistic assignment of a grade or use of the analytic rating scale--measures improvement over time, use of a holistically derived grade accounts for more of the difference in improvement over time.

FOOTNOTES

[1]Michael Clark, "Contests and Context: Writing and Testing in School," College English, November, 1980, p. 217.

[2]Anne Ruggles Gere, "Written Composition: Toward a Theory of Evaluation," College English, Vol. 42, No. 1, September 1980.

[3]Lee Odell and Charles Cooper, "Procedures for Evaluating Writing: Assumptions and Needed Research," College English, Vol. 42, No. 1, September 1980, pp. 35-43.

[4]Gertrude Conlan, "How the Essay in the CEEB English Test is Scored," (Urbana, Illinois: National Council of Teachers of English, 1976).

[5]Paul D. Diederich, Measuring Growth in English, (Urbana, Illinois: National Council of Teachers of English, 1974).

[6]E. D. Hirsch, Jr., The Philosophy of Composition, (Chicago: University of Chicago Press, 1977).

[7]Richard Lloyd-Jones, "Primary Trait Scoring," In Evaluating Writing: Describing, Measuring, Judging, Ed. Charles R. Cooper and Lee Odell (Urbana, Illinois: National Council of Teachers of English, 1977), pp. 33-66.

[8]Ibid.

[9]Lee Odell, "Teachers of Composition and Needed Research in Discourse Theory," College Composition and Communication, Vol. 30, No. 1, February 1979, pp. 39-45.

[10]Ibid.

[11]Charles R. Cooper, "Holistic Evaluation of Writing," in Evaluating Writing by Cooper and Odell. Urbana: NCTE, 1977, pp. 3-31.

[12]Peter Elbow, Writing Without Teachers. New York: Oxford University Press, 1973.

[13]Cooper, Op. Cit., p. 11.

[14]Conlan, 1976.

[15]Ruth Smith, Grading the Advanced Placement English Examination. Princeton, New Jersey: College Board Publications, 1975.

[16]Writing Mechanics, 1975.

[17]State Education Department, Hartford, Connecticut, 1980.

[18]Wayne Neuburger, "Writing Assessment in a Public School District," Clearing House for Applied Performance Testing Newsletter, Vol. 3, Issue 3, January 1978, p. 6.

[19]J. W. French, "Schools of Thought in Judging Excellence of English Themes." In A. Anastasi, Testing Problems in Perspective: 25th Anniversary Volume of Topical Readings from the Invitational Conference on Testing Problems. Washington, D.C., American Council on Education, 1966, pp. 587-596.

[20]Starch, D. and Elliott, E. C. "Reliability of Grading High School Work in English." School Review, 1912. 20, 442-457.

[21]H. C. Meckel, "Research on Teaching Composition and Literature," in N. L. Gage, Editor, Handbook of Research on Teaching. Chicago: Rand McNally, 1963, pp. 966-1066.

[22]Coffman, W. E. and Kurfman, D. "A Comparison of Two Methods of Reading Essay Examinations." American Educational Research Journal, 1968, 5, 99-107.

[23]Miriam Cherkes and Katherine Jones Loheyde. "The Cognitive Correlates of Analytic Prose." Unpublished paper presented at the annual meeting of the American Educational Research Association, Boston, April, 1980.

BIBLIOGRAPHY

Cherkes, Miriam and Jones Loheyde, Katherine. "The Cognitive Correlates of Analytic Prose." Unpublished paper presented at the annual meeting of the American Educational Research Association. Boston, April, 1980.

Clark, Michael. "Contests and Contexts: Writing and Testing in School." College English, November 1980, pp. 217-227.

Coffman, W. E. and Kurfman, D. "A Comparison of Two Methods of Reading Essay Examinations." American Educational Research Journal, 1968, 5, pp. 99-107.

Conlan, Gertrude. "How the Essay in the CEEB English Test is Scored." Urbana, Illinois: National Council of Teachers of English, 1976.

Cooper, Charles. "Holistic Evaluation of Writing." Evaluating Writing. Ed. C. Cooper and L. Odell. Urbana, Illinois: National Council of Teachers of English, 1977, pp. 3-31.

------ and Odell, Lee. Evaluating Writing: Describing, Measuring, Judging. Urbana, Illinois: National Council of Teachers of English, 1977.

------ and Odell, Lee. Research on Composing: Points of Departure. Urbana, Illinois: National Council of Teachers of English, 1978.

Diederich, Paul. Measuring Growth in English. Urbana, Illinois: National Council of Teachers of English, 1974.

Elbow, Peter. Writing Without Teachers. New York: Oxford University Press, 1973.

Foley, Joseph J. "Evaluation of Learning in Writing." Handbook of Formative and Summative Evaluation of Students Learning. Ed. Benjamin Bloom et. al. New York: McGraw-Hill, 1971.

French, J. W. "Schools of Thought in Judging Excellence of English Themes." Testing Problems in Perspective: 25th Anniversary Volume of Topical Readings from the Invitational Conference on Testing Problems. Ed. A. Anastasi. Washington, D. C.: American Council on Education, 1966, pp. 587-596.

Gere, Anne Ruggles. "Written Composition: Toward a Theory of Evaluation." College English, 1980, 42, (42) (1), 44.

Hirsch, E. D., Jr. The Philosophy of Composition. Chicago: University of Chicago Press, 1977.

Lloyd-Jones, Richard. "Primary Trait Scoring." in Evaluating Writing: Describing Measuring, Judging. Ed. C. Cooper and L. Odell. Urbana, Illinois: National Council of Teachers of English, 1977, pp. 33-66.

Meckel, H. C. "Research on Teaching Composition and Literature." Handbook of Research on Teaching. Ed. N. L. Gage. Chicago: Rand McNally, 1963, pp. 966-1066.

Neuburger, Wayne. "Writing Assessment in a Public School District." Clearinghouse for Applied Performance Testing Newsletter. Vol. 3, No. 3, January 1978, p. 6.

Odell, Lee. "Teachers of Composition and Needed Research in Discourse Theory." College Composition and Communication. Vol. 30, No. 1, February 1979, pp. 39-45.

------ and Cooper, Charles. "Procedures for Evaluating Writing: Assumptions and Needed Research." College English. Vol. 42, No. 1, September, 1980, pp. 35-43.

Smith, Ruth. Grading the Advanced Placement English Examination. Princeton, New Jersey: College Board Publications, 1975.

Starch, D. and Elliott, E. C. "Reliability of Grading High School Work in English." School Review. 20, 1912, pp. 442-457.

APPENDIX A

Readings to Which Students Responded

(1) "Women in Our Society" -- used to generate September paper.

(2) "A Proposal to Abolish Grading" -- used to generate December paper.

Women in our society complain of the lack of stimulation, of the loneliness, of the dullness of staying at home. Little babies are poor conversationalists, husbands come home tired and sit reading the paper, and women who used to pride themselves on their ability to talk find on the rare evening they can go out that their words clot on their tongues. As the children go to school, the mother is left to the companionship of the Frigidaire and the washing machine. Yet she can't go out because the delivery man might come, or a child might be sent home sick from school. The boredom of long hours of solitary one-sidedcommunication with things, no matter how shining and streamlined and new, descends upon her. Moreover, the conditions of modern life, apartment living, and especially the enormous amount of moving about, all serve to rob women of neighborhood ties. The better her electric equipment, the better she organizes her ordering, the less reason she has to run out for a bit of gossipy shopping at the corner store. The department stores and the moving-picture houses cater to women -- alone -- on their few hours out. Meanwhile efficient mending services and cheap ready-made clothes have taken most sensible busy work out of women's hands and left women -- still at home -- listening to the radio, watching television.

WHAT IS THE PROBLEM? HOW DO YOU THINK IT CAN BE SOLVED?

A PROPOSAL TO ABOLISH GRADING

Paul Goodman

Let half a dozen of the prestigious Universities--Chicago, Stanford, the Ivy League--abolish grading, and use testing only and entirely for pedagogic purposes as teachers see fit.

Anyone who knows the frantic temper of the present schools will understand the transvaluation of values that would be effected by this modest innovation. For most of the students, the competitive grade has come to be the essence. The naive teacher points to the beauty of the subject and the ingenuity of the research; the shrewd student asks if he is responsible for that on the final exam.

Let me at once dispose of an objection whose unanimity is quite fascinating. I think that the great majority of professors agree that grading hinders teaching and creates a bad spirit, going as far as cheating and plagiarizing. I have before me the collection of essays, Examining in Harvard College, and this is the consenus. It is uniformly asserted, however, that the grading is inevitable, for how else will the graduate schools, the foundations, the corporations know whom to accept, reward, hire? How will the talent scouts know whom to tap?

pedagogic - teaching

transvaluation - evaluating on a different basis

essence - primary characteristic

consensus - agreement in matters of opinion

WHAT IS THE PROBLEM? HOW DO YOU THINK IT CAN BE SOLVED?